ACOUSTICAL PRE-PROCESSING FOR ROBUST SPEECH RECOGNITION

Richard M. Stern and Alejandro Acero[1]
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

ABSTRACT

In this paper we describe our initial efforts to make SPHINX, the CMU continuous speech recognition system, environmentally robust. Our work has two major goals: to enable SPHINX to adapt to changes in microphone and acoustical environment, and to improve the performance of SPHINX when it is trained and tested using a desk-top microphone. This talk will describe some of our work in acoustical pre-processing techniques, specifically spectral normalization and spectral subtraction performed using an efficient pair of algorithms that operate primarily in the cepstral domain. The effects of these signal processing algorithms on the recognition accuracy of the Sphinx speech recognition system was compared using speech simultaneously recorded from two types of microphones: the standard close-talking Sennheiser HMD224 microphone and the desk-top Crown PZM6fs microphone. A naturally-elicited alphanumeric speech database was used. In initial results using the stereo alphanumeric database, we found that both the spectral subtraction and spectral normalization algorithms were able to provide very substantial improvements in recognition accuracy when the system was trained on the close-talking microphone and tested on the desk-top microphone, or vice versa. Improving the recognition accuracy of the system when trained and tested on the desk-top microphone remains a difficult problem requiring more sophisticated noise suppression techniques.

INTRODUCTION

The acceptability of any voice interface depends on its ease of use. Although users in some application domains will accept the headset-mounted microphones that are commonly used with current speech recognition systems, there are many other applications that require a desk microphone or a wall-mounted microphone. The use of other types of microphones besides the "close-talking" headset generally degrades the performance of spoken-language systems. Even a relatively "quiet" office environment can be expected to provide a significant amount of additive noise from fans, door slams, as well as competing conversations and reverberation arising from surface reflections within a room. Applications such as inspection or inventory on a factory floor, or an outdoor automatic banking machine demand an even greater degree of environmental robustness. Our goal has been to develop practical spoken-language systems for real-world environments that are robust with respect to changes in acoustical ambience and microphone type as well as with respect to speaker and dialect.

Although a number of techniques have been proposed to improve the quality of degraded speech, researchers have only recently begun to evaluate speech-enhancement in terms of the improvement in recognition accuracy that they provide for speech-recognition systems operating in natural environments. We are incorporating into our system a combination of techniques that come into play at different levels of the system, including pre-processing of the acoustical waveform, the development of physiologically and psychophysically motivated peripheral processing models (*i.e.* "ear models"), adaptive multimicrophone array processing, and dynamic adaptation to new speakers and environments by modifying the parameters used to represent the speech sounds. In this talk we will focus only on our work in the first category, acoustical preprocessing.

---

[1]This research was sponsored by the Defense Advanced Research Projects Agency (DOD), ARPA Order No. 5167, under contract number N00039-85-C-0163. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Defense Advanced Research Projects Agency or the US Government.

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
1989		**00-00-1989 to 00-00-1989**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Acoustical Pre-Processing for Robust Speech Recognition	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Carnegie Mellon University,School of Computer Science,Pittsburgh,PA,15213**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**		**8**	

There are many sources of acoustical distortion that can degrade the accuracy of speech-recognition systems. For example, obstacles to robustness include additive noise from machinery, competing talkers, etc., reverberation from surface reflections in a room, and spectral shaping by microphones and the vocal tracts of individual speakers. These sources of distortion cluster into two complementary classes: *additive* noise (as in the first two examples) and distortions resulting the *convolution* of the speech signal with an unknown linear system (as in the remaining three).

In the classical speech-enhancement literature, two complementary techniques have been proposed to cope with these problems: spectral subtraction and spectral normalization. In spectral subtraction one estimates the amount of background noise present during non-speech intervals, and *subtracts* the estimated spectral density of the noise from the incoming signal (*e.g.* Boll, 1979; Berouti *et al.*, 1979). In spectral normalization (sometimes referred to as "blind deconvolution"), one estimates the average spectrum when speech is present and applies a *multiplicative* normalization factor with respect to a reference spectrum (*e.g.* Stockham *et al.*, 1975). While these procedures were once thought to be of limited practical benefit, based on the results of experiments concerning the human perception of speech, results of recent applications of them to automatic speech-recognitions systems have been more encouraging (*e.g.* Porter and Boll, 1984; Van Compernolle, 1987).

In this report we will review the database used to evaluate efficient implementations of spectral subtraction and normalization in the cepstral domain, discuss the results of analyses of baseline studies of recognition performance, describe the effectiveness of the spectral subtraction and normalization algorithms, and discuss the motivations for some of our work in progress.

THE ALPHANUMERIC DATABASE

Although the bulk of research using the Sphinx system at Carnegie Mellon has made use of the well-known Resource Management database, we were forced to use a different database, the Alphanumeric database, for our evaluations of signal processing. The primary reason for this is that the Resource Management database with its large vocabulary size and many utterances required several weeks to train satisfactorily, which was excessively long since the entire system had to be retrained each time a new signal-processing algorithm was introduced. We also performed these evaluations using a more compact and easily-trained version of Sphinx with only about 650 phonetic models, omitting such features as function-word models, between-word triphone models, and corrective training. We were willing to tolerate the somewhat lower absolute recognition accuracy that this version of Sphinx provided because of the reduced time required by the training process. Using the Alphanumeric database, the more compact Sphinx system, and faster computers, we were able to reduce the training time to the point that an entire train-and-test cycle could be performed in about 9 hours.

A second reason why we resorted to a new database is that we specifically wanted to compare simultaneous recordings from close-talking and desktop microphones in our evaluations. We believe that it is very important to evaluate speech-recognition systems in the context of natural acoustical environments with natural noise sources, rather than using speech that is recorded in a quiet environment into which additive noise and spectral tilt are artificially injected.

CONTENTS OF THE DATABASE

The Alphanumeric database consists of 1000 training utterances and 140 different testing utterances, that were each recorded simultaneously in stereo using both the Sennheiser HMD224 close-talking microphone that has been a standard in previous DARPA evaluations, and a desk-top Crown PZM6fs microphone. The recordings were made in one of the CMU speech laboratories (the "Agora" lab), which has high ceilings, concrete-block walls, and a carpeted floor. Although the recordings were made behind an acoustic partition, no attempt was made to silence other users of the room during recording sessions, and there is consequently a significant amount of audible interference from other talkers, key clicks from other workstations, slamming doors, and other sources of interference, as well as the reverberation from the room itself. Since the database was limited in size, it was necessary to perform repeated evaluations on the same test utterances.

The database consisted of strings of letters, numbers, and a few control words, that were naturally elicited in the context of a task in which speakers spelled their names, addresses, and other personal information, and entered some random letter and digit strings. Some sample utterances are **N-S-V-H-6-T-49**, **ENTER-4-5-8-2-1** and

P-I-T-T-S-B-U-R-G-H. A total of 106 vocabulary items appeared in the vocabulary, of which about 40 were rarely uttered. Although it contains fewer vocabulary items, the Alphanumeric database is more difficult than the Resource Management database with perplexity 60 both because of the greater number of words in the vocabulary and because of their greater intrinsic acoustic confusibility.

AVERAGE SPEECH AND NOISE SPECTRA

Figure 1 compares averaged spectra from the Alphanumeric database for frames believed to contain speech and background noise from each of the two microphones. By comparing these curves, it can be seen that the average signal-to-noise ratio (SNR) using the close-talking Sennheiser microphone is about 25 dB. The signals from the Crown PZM, on the other hand, exhibit an SNR of less than 10 dB for frequencies below 1500 Hz and about 15 dB for frequencies above 2000 Hz. Furthermore, the response of the Crown PZM exhibits a greater spectral tilt than that of the Sennheiser, perhaps because the noise-cancelling transducer on the Sennheiser also suppresses much of the low-frequency components of the speech signal.

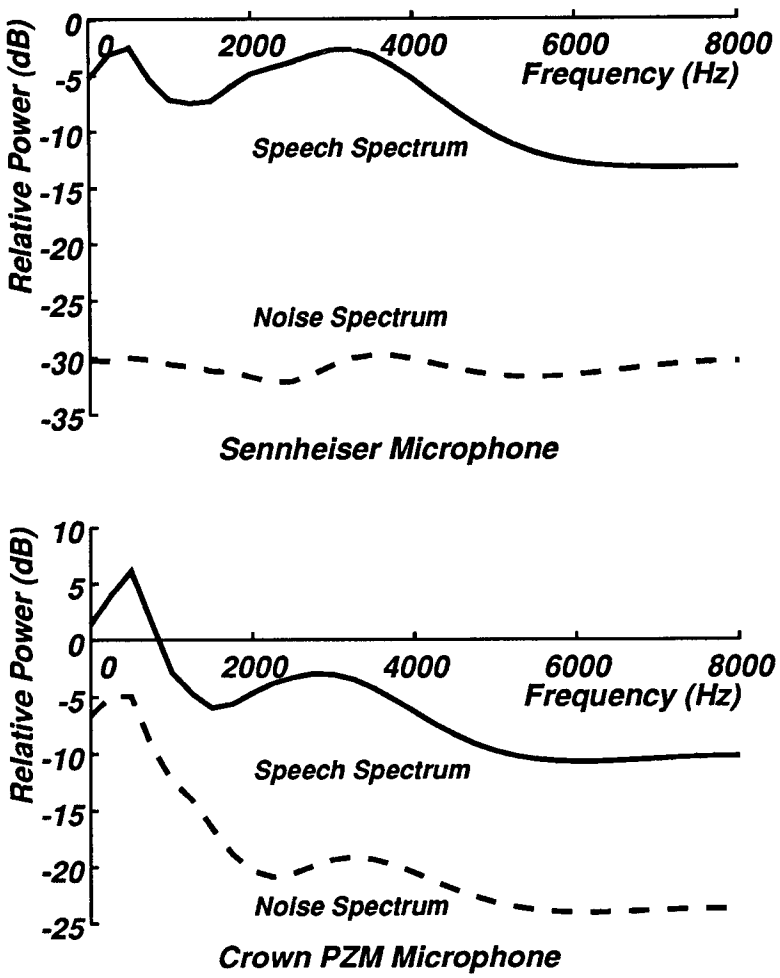

Figure 1 Average speech and noise spectra from the Alphanumeric database obtained using the headset-mounted Sennheiser Microphone and the Crown PZM microphone. The separation of the two curves in each panel provides an indication of signal-to-noise ratio for each microphone. It can also be seen that the Crown PZM produces greater spectral tilt.

BASELINE RECOGNITION ACCURACY

We first consider the "baseline" recognition accuracy of the Sphinx system obtained using the two microphones with the standard signal processing routines. Table I summarizes the recognition accuracy obtained by training and testing using each of the two microphones. Recognition accuracy is reported using the standard DARPA scoring procedure (Pallett, 1989), with penalties for insertions and deletions as well as for substitutions. It can be seen that training and testing on the Crown PZM produces an error rate that is 60% worse than the error rate produced when the system is trained and tested on the Sennheiser microphone. When the system is trained using one microphone and tested using the other, however, the performance degrades to a very low level. Hence we can identify two goals of signal processing for greater robustness: we need to drastically improve the performance of the system for the "cross conditions", and to elevate the absolute performance of the system when it is trained and tested using the Crown PZM.

	Test CLS	Test PZM
Train CLS	85.3 %	18.6%
Train PZM	36.9%	76.5%

Table I. Baseline performance of the Sphinx system when trained and tested on the Alphanumeric vocabulary using each of the two microphones.

In order to better understand why performance degraded when the microphone was changed from the Sennheiser to the Crown PZM, even when the PZM was used for training as well as testing, we studied the spectrograms and listened carefully to all utterances for which training and testing with the PZM produced errors that did not appear when the system was trained and tested on the close-talking Sennheiser microphone. The estimated causes of the "new" errors using the Crown PZM are summarized in Table II. Not too surprisingly, the major consequence of using the PZM was that the effective SNR was lowered. As a result, there were many confusions of silence or noise segments with weak phonetic events. These confusions accounted for some 58 percent of the additional errors, with crosstalk (either by competing speakers or key clicks from other workstations) identified as the most significant other cause of new errors.

	Percent errors
Weak-event insertion	41.5
Weak-event deletion	13.2
Crosstalk	20.0
Others	25.3

Table II. Analysis of causes of "new" errors introduced by use of the Crown PZM microphone.

We now consider the extent to which the use of acoustical pre-processing can mitigate the effects of the Crown PZM and of the change in microphone.

ACOUSTICAL PRE-PROCESSING FOR SPEECH RECOGNITION

In this section we briefly review the baseline signal procedures used in the Sphinx system, and we describe the spectral normalization and spectral subtraction operations in the cepstral domain.

GENERAL SIGNAL PROCESSING

The first stages of signal processing in the evaluation system are virtually identical to those that have been reported for the Sphinx system previously. Briefly, speech is digitized with a sampling rate of 16 kHz and pre-emphasized, and a Hamming window is applied to produce analysis frames of 20-ms duration every 10 ms. 14 LPC coefficients

are produced for each frame using the autocorrelation method, from which 32 cepstral coefficients are obtained using the standard recursion method. Finally, these cepstral coefficients are frequency warped to a pseudo-mel scale using the bilinear-transform method with 12 stages, producing a final 12 cepstral coefficients after the frequency warping. (We found that increasing the number of cepstral coefficients before the warping from 12 to 32 provided better frequency resolution after frequency warping, which led to a 5-percent relative improvement of the baseline Sphinx system on the Resource Management task.) In addition to the LPC cepstral coefficients, differenced LPC cepstral coefficients, power and differenced power are also computed for every frame. The cepstra, differenced cepstra, and combined power and differenced power parameters are vector quantized into three different codebooks.

PROCESSING FOR ROBUSTNESS IN THE CEPSTRAL DOMAIN

We describe in this section the procedures we used to achieve spectral normalization and spectral subtraction in the cepstral domain. Because signal processing and feature extraction in the Sphinx system was already based on cepstral analysis, these procedures could be implemented with an almost negligible increase in computational load beyond that of the existing signal processing procedures.

Spectral Normalization

The goal of spectral normalization is to compensate for distortions to the speech signal produced by linear convolution, which could be the result of filtering by the vocal tract, room acoustics, or the transfer function of a particular microphone. As noted above, compensation for linear convolution could be accomplished by multiplying the magnitude of the spectrum by a correction factor. Since the cepstrum is the log of the magnitude of the spectrum, this corresponds to a simple additive correction of the cepstrum vector. The major differences between various spectral normalization algorithms are primarily concerned with how the additive compensation vector is estimated.

The most effective form of spectral normalization that we have considered so far is also the simplest. Specifically, a *static* reference vector is estimated by computing the inverse DFT of the long-term average of the cepstral vector for the speech frames from the training databases. (Samples of these averages for the alphanumeric database are shown in Fig. 1.) The compensation vector is defined to be the difference between the two sets of averaged cepstral coefficients from the two types of microphones in the training database. Although the compensation vector is determined only from averages of spectra in the speech frames, it is applied to both the speech and nonspeech frames.

We have also considered other types of spectral normalization in the cepstral domain, including one that determines the compensation vector that minimizes the average VQ distortion. While none of these methods work any better in isolation than the simple static spectral normalization described above, some of them have exhibited better performance than the static normalization when used in conjunction with spectral subtraction.

Spectral Subtraction

Spectral Subtraction is more complex than spectral normalization, both because it cannot be applied to the cepstral coefficients directly, and because there are more free parameters and arbitrary decisions that must be resolved in determining the best procedure for a particular system.

Spectral subtraction in the Sphinx system is accomplished by converting from the feature vectors from cepstral coefficients to log-magnitude coefficients using a 32-point inverse DFT (for the 16 real and even cepstral coefficients). These log-magnitude vectors are then exponentiated to produce direct spectral magnitudes, from which a reference vector is subtracted according to the general procedure described below. The log of the resulting difference spectrum is then converted once again to a cepstral vector using a 32-point forward DFT. Although both an inverse and forward DFT must be performed on the cepstral vectors in this algorithm, little time is consumed because only 16 real coefficients are involved in the DFT computations. In addition, a computationally efficient procedure similar to the one described by Von Compernolle (1987) can be applied to perform the exponentiation and logarithm operations using a single table lookup.

The estimated noise spectrum is either over-subtracted or under-subtracted from the input spectrum, depending on the estimated instantaneous signal-to-noise ratio (of the current analysis frame). In our current implementation of

spectral subtraction, the estimation of the noise vector and the determination of the amount of subtraction to be invoked are based on a comparison of the incoming signal energy to two thresholds, representing a putative maximum power level for noise frames (the "noise threshold") and a putative minimum power level for speech frames (the "speech" threshold"). While these thresholds are presently set empirically, they could easily be estimated from histograms of the average power for the signals in the analysis frames. The estimated noise vector is obtained by averaging the cepstra of all frames with a power that falls below the noise threshold. Once the noise vector is estimated, a magnitude equal to that of the reference spectrum plus 5 dB is subtracted from the magnitude of the spectrum of the incoming signal, for all frames in which the power of the incoming signal falls below the noise threshold. If the power of the incoming signal is above the speech threshold, the magnitude of the reference spectrum *minus* 2.5 dB is subtracted from the magnitude of the spectrum of the incoming signal. The amount of over- or under-subtraction (in dB) is a linearly interpolated function of the instantaneous signal-to-noise ratio (in dB) for incoming signals whose power is between the two thresholds. We note that we subtract the magnitudes of spectra [as did Berouti *et al.* (1979)] rather than the more intuitively appealing spectral power because we found that magnitude subtraction provides greater recognition accuracy.

EXPERIMENTAL RESULTS

Figure 2 summarizes the experimental results obtained using the Alphanumeric database when the system was trained and tested on the two types of microphones, in either the baseline conditions, or with spectral normalization and spectral subtraction. In each of the two panels, the word accuracies obtained for the two baseline conditions when the system was trained and tested using the same microphones are indicated by the horizontal dotted lines. It can be seen that in each case, the use of spectral normalization and subtraction provides increasing improvement to the recognition accuracy obtained in the "cross" conditions, without almost no degradation of the recognition accuracy observed when the system is trained and tested using the same microphone. In fact, the recognition accuracy obtained with spectral subtraction in the "cross" conditions approaches that obtained when the system is trained on the same microphone that it is tested on. On the other hand, we have not yet been able to significantly improve the performance of the system when it is trained and tested on the Crown PZM microphone. We briefly describe some of the strategies we are presently considering toward that end.

DISCUSSION

We demonstrated in the previous section that the spectral subtraction and normalization routines we have implemented can greatly increase the robustness of the Sphinx system when it is tested on a different microphone from the one with which it was trained. While we are pleased with these results, we are also continuing our efforts to improve the performance of the system when trained and tested using the Crown PZM microphone. We strongly believe that further improvements in performance are possible for this condition using improved acoustical pre-processing, and we briefly describe three techniques to be considered.

INTEGRATION OF SPECTRAL SUBTRACTION AND NORMALIZATION

Since spectral subtraction and normalization each provide some improvement in recognition accuracy when applied individually, one would expect that further improvement should be obtained when they are used simultaneously. Indeed, in pilot experiments using the Resource Management database, training using the Sennheiser microphone and testing using the Crown PZM, we obtained a 15 percent reduction in relative error rate when spectral normalization was added to spectral subtraction (Morii, 1987). Nevertheless, we have found that the effects of the two enhancement procedures interact with each other, and simple cascades of the two implementations that work best in isolation do not produce great improvements in performance. We are confident that with better understanding of the nature of these interactions we can more fully exploit the complementary nature of the two types of processing.

INTRODUCTION OF NON-PHONETIC MODELS

In these Proceedings, Ward (1989) describes a procedure by which the performance of the Sphinx system can be improved by explicitly developing phonetic models for such non-speech events such as filled pauses, breath noises, door slams, telephone rings, paper rustling, etc. Most of these phenomena are highly transitory in nature, and as such are not directly addressed by either spectral subtraction or normalization. While Ward was especially

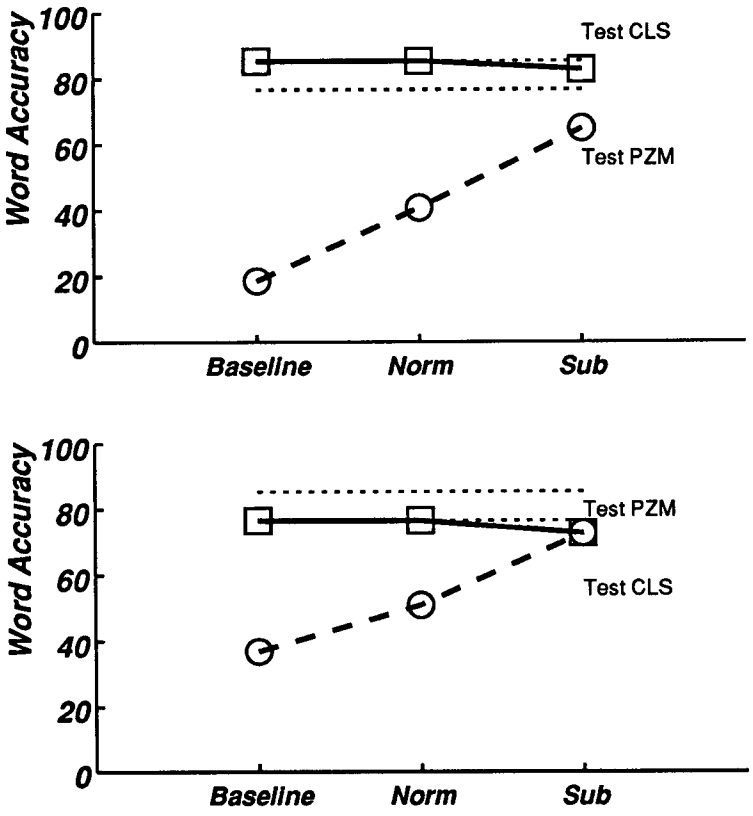

Figure 2 Comparison of recognition accuracy obtained using the baseline signal processing, spectral subtraction, and spectral normalization, and each of the two microphones. The horizontal dotted lines indicate performance obtained in the baseline condition when the system is trained and tested using the same microphone.

concerned with the non-phonetic events associated with spontaneous speech, there is no reason why these techniques cannot be applied to process speech recorded from desk-top microphones as well. Since it appears that about 20 percent of the "new" errors introduced when one replaces the Sennheiser microphone by the Crown PZM are the result of crosstalk, we are optimistic that implementation of Ward's non-phonetic models should provide further improvement in recognition accuracy.

CONSIDERATION OF SPECTRAL CORRELATIONS ACROSS FREQUENCY

Traditional spectral subtraction techniques assume that all speech frames are statistically independent from each other, and that every frequency component within a frame is statistically independent from the other frequencies. As a result, it is quite possible that the result of a spectral subtraction operation may bear little resemblance to any legitimate speech spectrum, particularly at low SNRs. We are exploring several techniques to take advantage of information about correlations across frequency to ensure that the result of the spectral subtraction is likely to represent a legitimate speech spectrum.

SUMMARY

We found that the use of desk-top microphones like the Crown PZM increase the error rate by allowing weak phonetic events to become confused with silences and vice-versa. The spectral subtraction and normalization routines we developed provide considerable improvement in recognition accuracy when the system is tested using a different microphone from the one it was trained on, but further work must be done to improve the absolute level of performance obtained when Sphinx is trained and tested using the Crown PZM.

ACKNOWLEDGMENTS

Many members of the speech group have contributed to this work. We thank Joel Douglas for performing many of the calculations, Kai-Fu Lee for helping us understand the mysteries of Sphinx, Fil Alleva and Eric Thayer for many discussions about signal processing in the Sphinx system, Bob Weide for providing for database collection and analysis, Wayne Ward for working with us to introduce non-phonetic models, and (of course) Raj Reddy for his overall leadership and support of this work.

REFERENCES

M. Berouti, R. Schwartz and J. Makhoul. (1979). Enhancement of Speech Corrupted by Acoustic Noise. In J. S. Lim (Ed.), *Speech Enhancement*. Englewood Cliffs, NJ: Prentice Hall, 1983.

S. F. Boll. (1979). Suppression of Acoustic Noise in Speech Using Spectral Subtraction. *ASSP*, *27*, 113-120.

S. Morii. (1987). Performance of the Sphinx System Using Spectral Subtraction and Normalization. Unpublished work, Carnegie Mellon University.

D. Pallett. (1989). Benchmark Tests for DARPA Resource Management Database Performance Evaluations. *ICASSP89*.

J. E. Porter and S. F. Boll. (1984). Optimal Estimators for Spectral Restoration of Noisy Speech. *ICASSP84*.

T. G. Stockham, T. M. Cannon and R. B. Ingebretsen. (1975). Blind Deconvolution Through Digital Signal Processing. *Proc. IEEE*, *63*, 678-692.

D. Van Compernolle. (1987). Increased Noise Immunity in Large Vocabulary Speech Recognition with the Aid of Spectral Subtraction. *ICASSP87*.

W. Ward. (1989). Modelling Non-Verbal Sounds for Speech Recognition. These Proceedings.